Globus Online for APS Data Movement
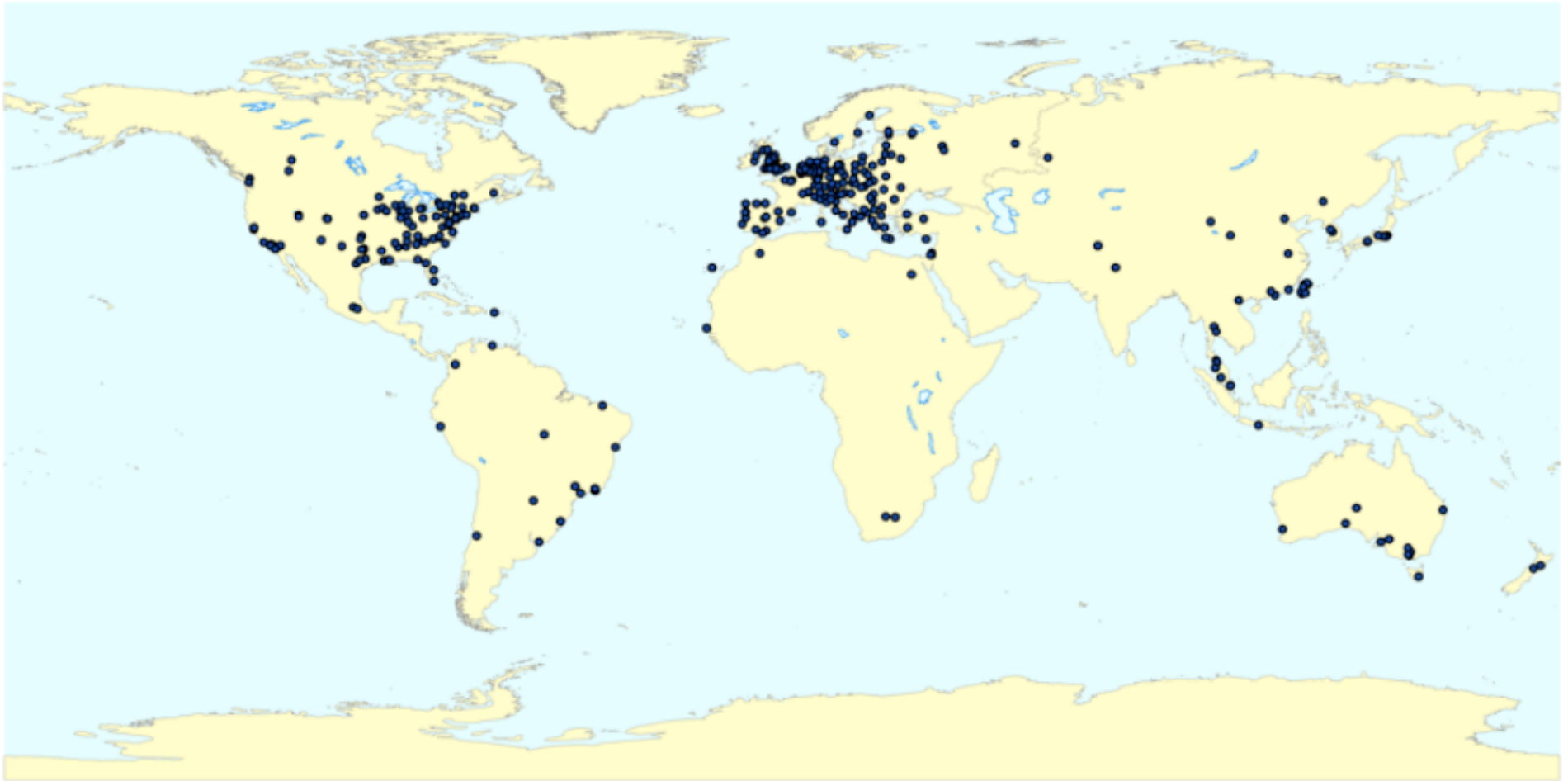
Raj Kettimuthu

5/3/2011

GridFTP

- **High-performance, reliable data transfer protocol optimized for high-bandwidth wide-area networks**

- **Globus GridFTP**

  - High Performance

    - Parallel TCP streams, Non TCP protocol such as UDT, Cluster-to-cluster data movement

  - Secure
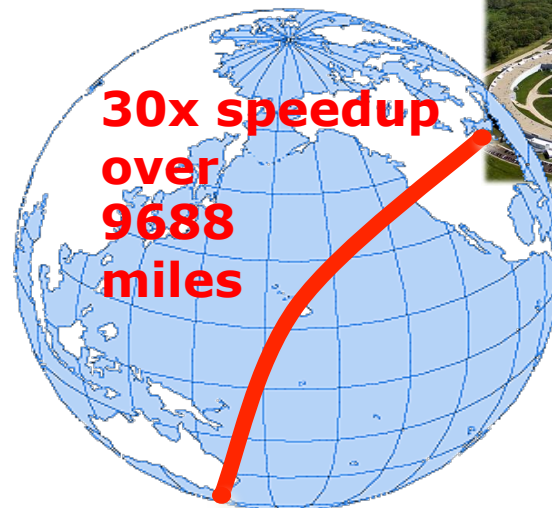
  - Synchronize data sets

  - Checkpointing

APS

GridFTP

Created by Tim Pinkawa (Northern Illinois University) using MaxMind's GeoIP technology (http://www.maxmind.com/app/ip-locate).

APS

30x speedup over 9688 miles

One terabyte moved from an Advanced Photon Source to Australia, at a rate 30x faster than SCP

Discover endpoints, determine available protocols, negotiate firewalls, configure software, manage space, determine required credentials, configure protocols, detect and respond to failures, determine expected performance, determines actual performance, identify diagnose and correct network misconfigurations, integrate with file systems, …

A → B

APS
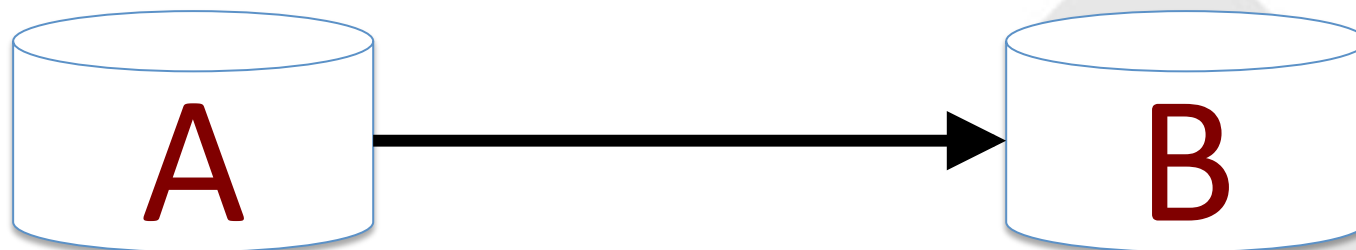
Globus Online

- **Easy "fire and forget" file transfers**
- **Automatic fault recovery**
- **High performance**
- **Simplify use of multiple security domains**
- **No client software installation**
- **New features automatically available**
- **Consolidated support and troubleshooting**

APS
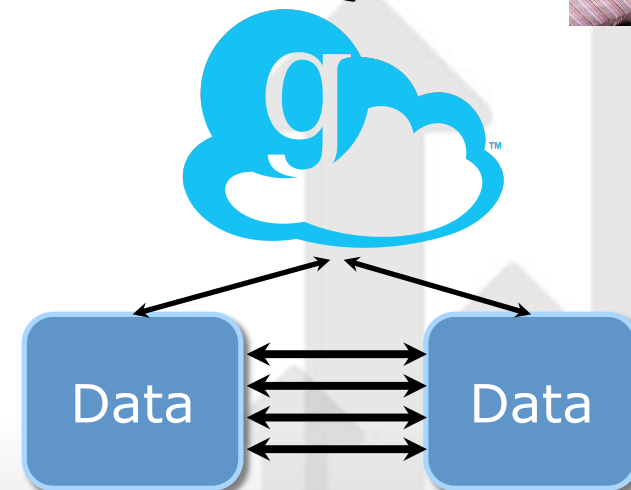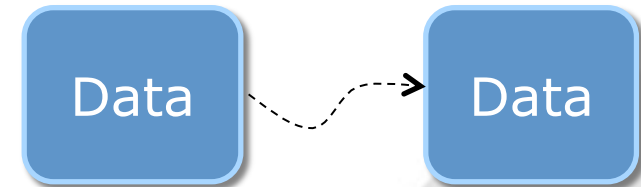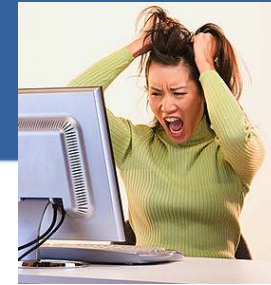
Data Movement Scenarios

- **Transfer between two GridFTP server**
  - Error recovery
  - Auto tuning

- **Upload/download data to/from a GridFTP server to/ from a user's machine**
  - Typically no GridFTP server on user's machine
  - GridFTP installation and security configuration
  - Globus Connect

- **Transfer between endpoints – neither has GridFTP server running**
  - Globus Connect for departments

globus online
Reliable File Transfer. No IT Required.

Reliable, high-performance, secure file transfer
Move files fast. No IT required.

+ HOW IT WORKS
Learn more about the service

> GET STARTED
Setup profile in 3 easy steps

Globus Online makes robust file transfer capabilities, traditionally available only on expensive, special-purpose software systems, accessible to everyone.

Learn more

Why Use Globus Online?
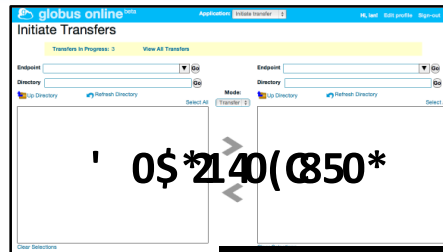See how easy file transfer can

For HPC Resource Owners
Enable Globus Online

For Developers
Integrate with Globus Online

Globus Online

Globus Connect

Globus Connect Installation

Globus Connect allows you to use Globus Online to transfer files to and from your computer.

Need Help? Click Here

Step One: Choose Your Download

- Globus Connect For Mac OS X
- Globus Connect For Linux
- Globus Connect For Windows

Step Two: Get Your Globus Connect Setup Key

Endpoint Name: raj-laptop

Description: (optional) Enter a description for this new endpoint

Generate Setup Key

Step Three: Finish Globus Connect Setup

Copy the setup key displayed above. Run Globus Connect and paste the key into the Initial Setup window when prompted. This setup key can only be used once.

Close

Globus Connect
1 of 2 selected, 90.4 MB available

globus online
Reliable File Transfer. No IT Required.

Globus Connect → Applications
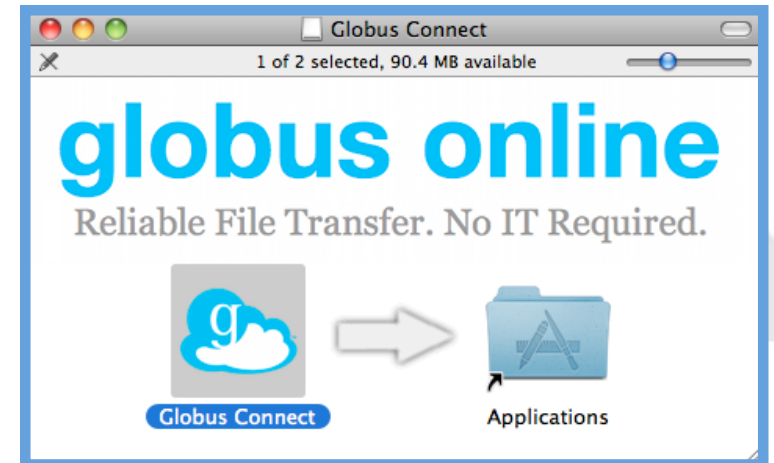
Setup

Initial Setup

Please type or paste your Globus Connect setup key into the field below and click 'OK' when finished.

Setup Key: 432e8ba5-45cf-442b-a374-5a8d1cfa75cb

▶ Advanced

Ok

Globus Connect

- **http://confluence.globus.org/display/GOPUB/Globus +Connect**

Globus Connect in Action

User

(2) User makes request Globus Online: e.g.,"transfer data from MyDesktop to SiteA"

Globus Online

1) Globus Connect client registers with Globus Online

(3) Globus Online forwards requests to Globus Connect & Server at Site A

GridFTP server "SiteA"

Globus Connect

4) Globus Connect establishes data channel connection to SiteA and transfers data

"MyDesktop"

Globus Connect in Action

User

(2) User makes request
Globus Online: e.g.,"transfer
data from MyDesktop to SiteA

Globus Online

1) Globus Connect
client registers with
Globus Online

3) Globus Online
forwards requests
to Globus Connect
& Server at Site A

GridFTP
server
"SiteA"

Globus
Connect

4) Globus Connect establishes data channel
connection to SiteA and transfers data

"MyDesktop"

Globus Connect in Action

User

(2) User makes request Globus Online: e.g.,"transfer data from MyDesktop to SiteA"

Globus Online

(1) Globus Connect client registers with Globus Online

(3) Globus Online forwards requests to Globus Connect & Server at Site A

GridFTP server "SiteA"

(4) Globus Connect establishes data channel connection to SiteA and transfers data

Globus Connect

"MyDesktop"

Globus Connect in Action

User

GridFTP server "SiteA"

Globus Online

Globus Connect

"MyDesktop"

(2) User makes request Globus Online: e.g.,"transfer data from MyDesktop to SiteA"

1) Globus Connect client registers with Globus Online

(3) Globus Online forwards requests to Globus Connect & Server at Site A

4) Globus Connect establishes data channel connection to SiteA and transfers data

Globus Online

GridFTP

GridFTP

Notification target

Profiles + state

Worker

User Gateway

amazon.com web services

APS

Globus Online in Action

28.6 Terabytes
31,000 files
56h 44m
No human involvement
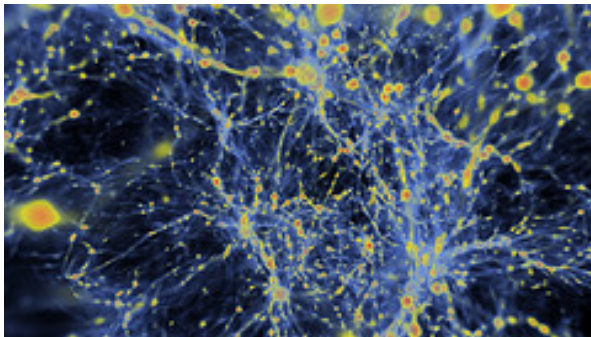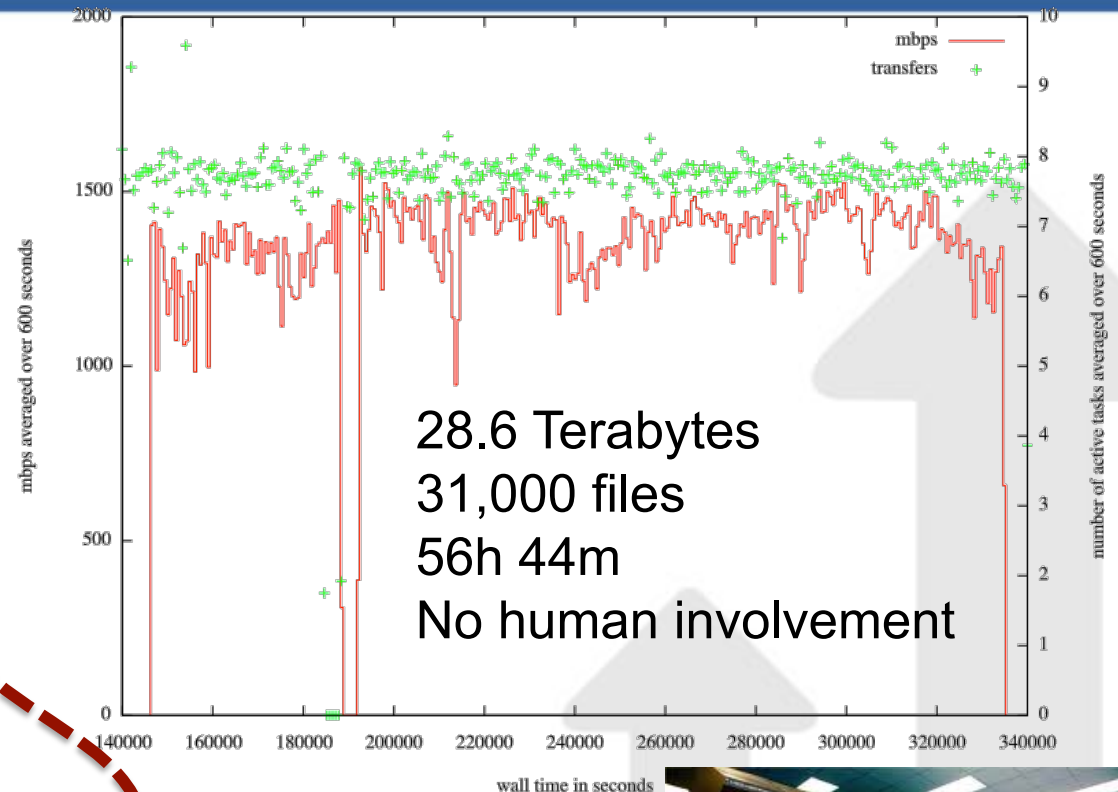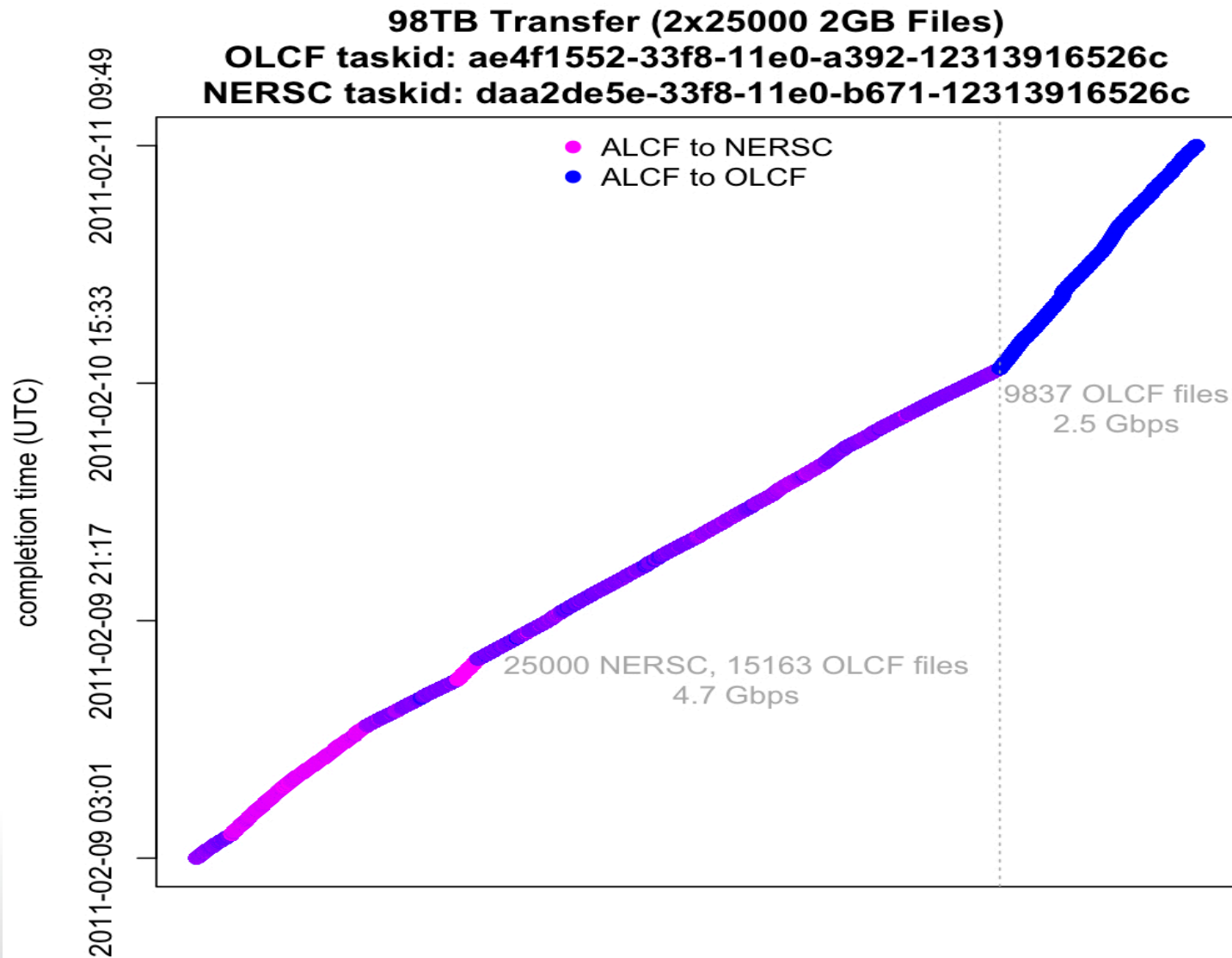
Astrophysics simulation data
generated in Tennessee,
moved to Illinois for visualization
(Enzo, UCSD; Futures Lab, Argonne)

APS

98TB Transfer (2x25000 2GB Files)
OLCF taskid: ae4f1552-33f8-11e0-a392-12313916526c
NERSC taskid: daa2de5e-33f8-11e0-b671-12313916526c

APS

Auto tuning

- **Tuning GridFTP to maximize performance is tricky**
- **Depends on the characteristics of the data sets**
  - Large files
  - Lots of small files
- **Depends on the characteristics of the network**
  - Dedicated network
  - Shared network
- **Globus Online does auto tuning**
  - Groups files into different buckets
 - Provides manual control for advanced users

Sync feature

- **Feature to synchronize datasets**
- **Only transfer files where the destination does not exist or differs (size, timestamp, checksum) from the source**
- **'-sync-level'**
  - 0 - transfer if the destination does not exist
  - 1 - transfer if the size does not match
  - 2 - transfer if timestamp of dst is older
  - 3 - transfer if the checksums do not match.
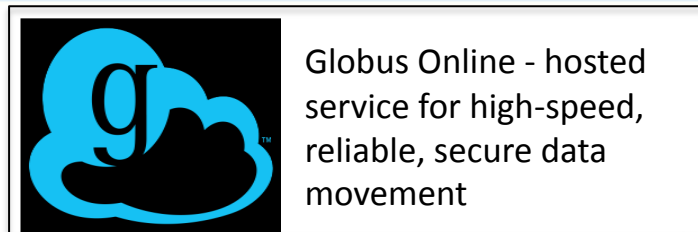 - The default sync level is 2.

Globus Online for APS Users
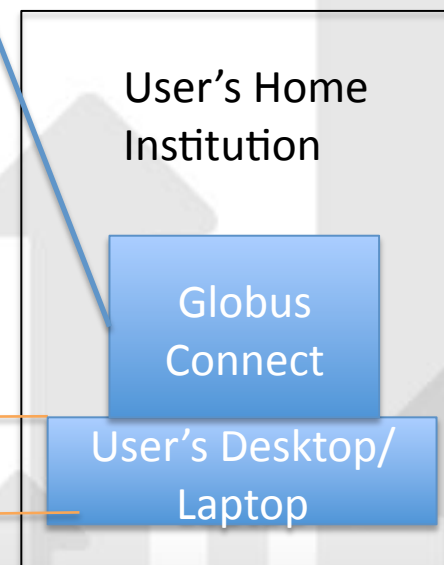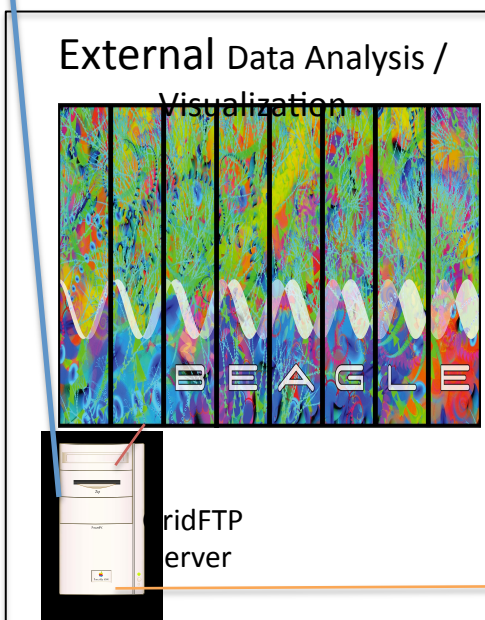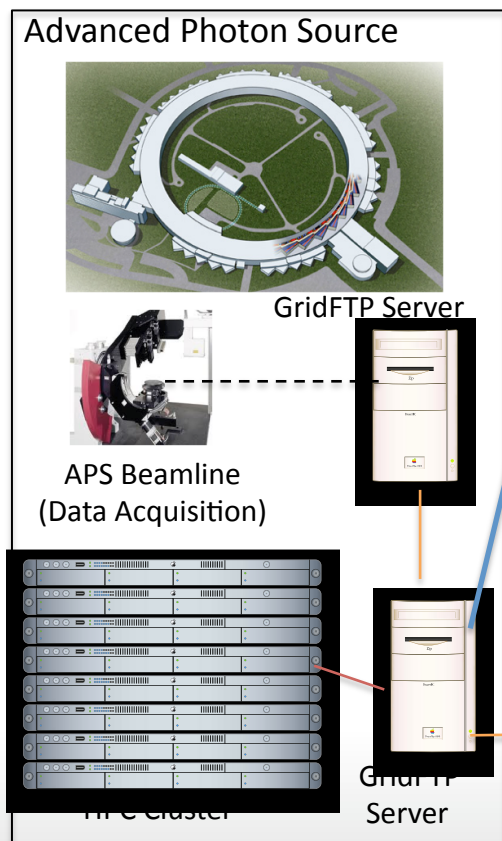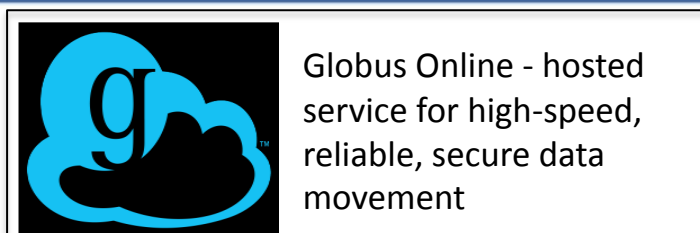
Globus Online - hosted service for high-speed, reliable, secure data movement
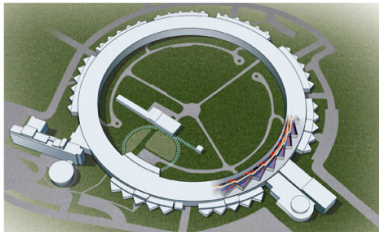
Advanced Photon Source

GridFTP Server

APS Beamline
(Data Acquisition)

HPC Cluster

GridFTP Server

APS

External Data Analysis / Visualization

BEAGLE

GridFTP Server

User's Home Institution

Globus Connect

User's Desktop/ Laptop

Globus Online for APS Users

Globus Online - hosted service for high-speed, reliable, secure data movement

Advanced Photon Source

GridFTP Server

APS Beamline (Data Acquisition)

Myproxy CA

HPC Cluster

GridFTP Server

External Data Analysis / Visualization
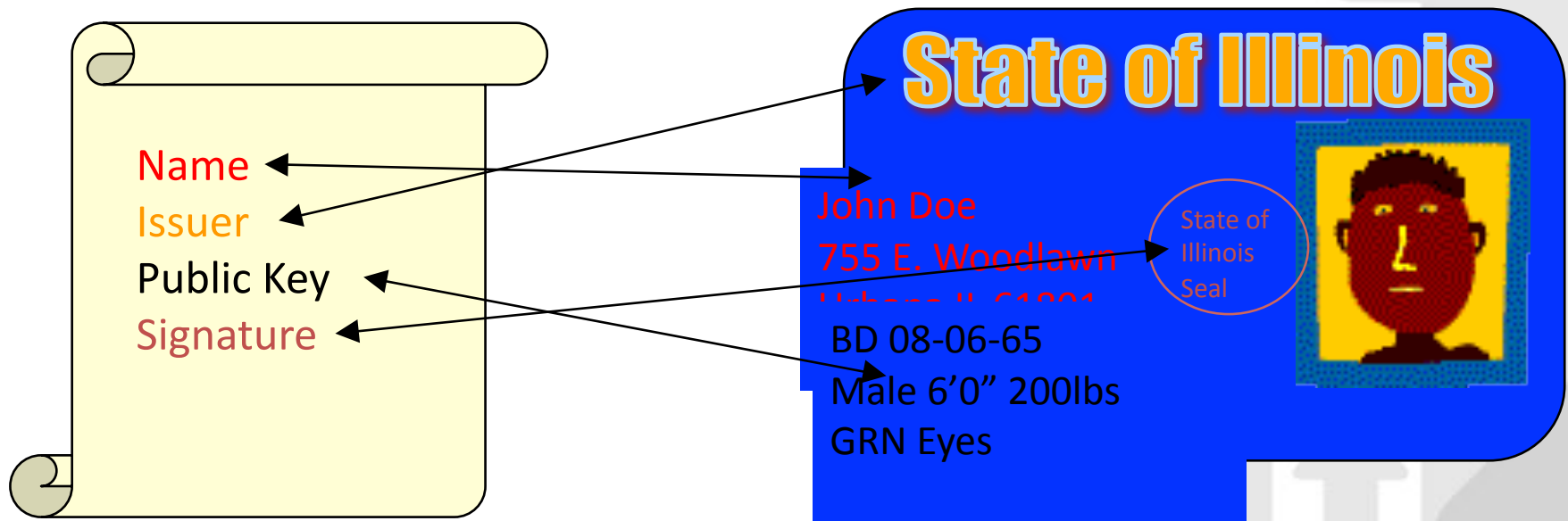
BEAGLE

GridFTP Server

User's Home Institution

Globus Connect

User's Desktop/ Laptop

APS

Certificates

- **Digital document that certifies a public key is owned by a particular user**
- **Signed by 3rd party – the Certificate Authority (CA)**
- **X509 standard**

Name
Issuer
Public Key
Signature

State of Illinois

John Doe
755 E. Woodlawn
Urbana, IL 61801

State of Illinois Seal

BD 08-06-65
Male 6'0" 200lbs
GRN Eyes

To know if you should trust the certificate, you have to trust the CA

APS

GSI Authentication

- **Public key based security**
  - X509 certificates
  - Short lived proxy credentials

- **Delegation**
  - Authenticates DC on clients behalf
  - Flexibility for automated workflows
    - Agents can authenticate to GridFTP servers on users behalf
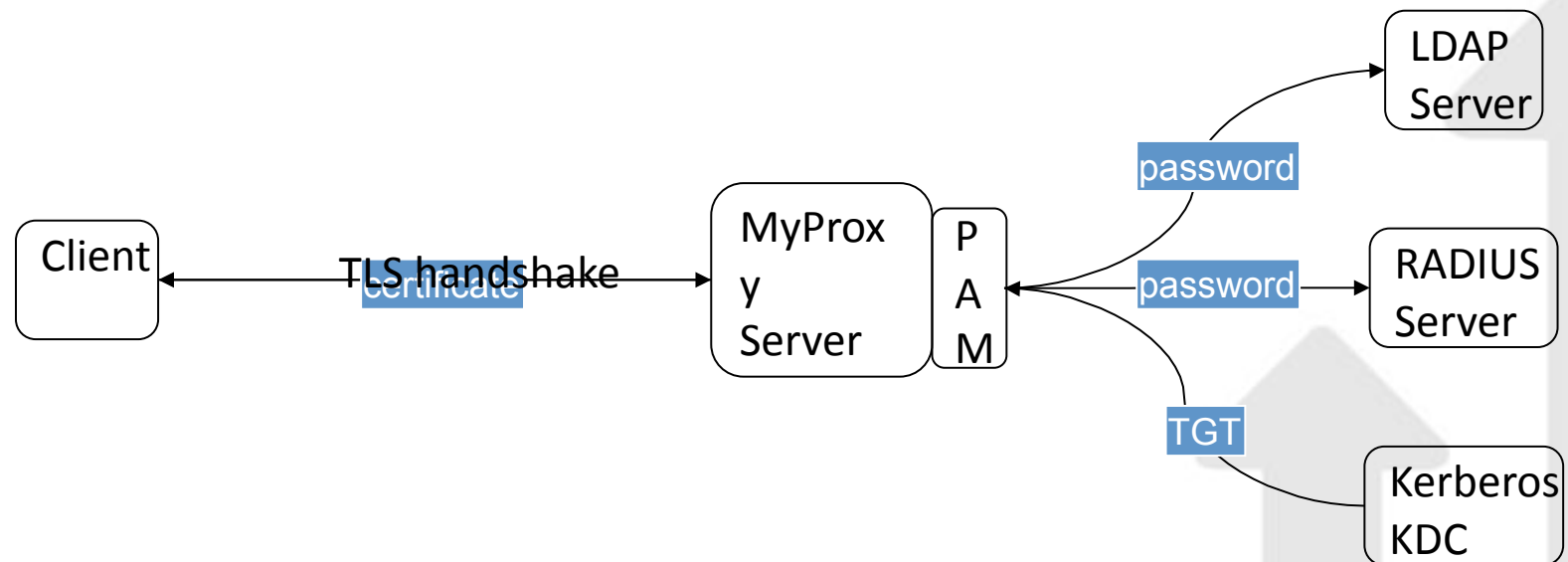 - Enables encryption, integrity on data channel

APS

GSI Authentication

GSI Credential

GridFTP Server A

Client

GSI Delegated
Credential

GSI Credential

GridFTP Server B
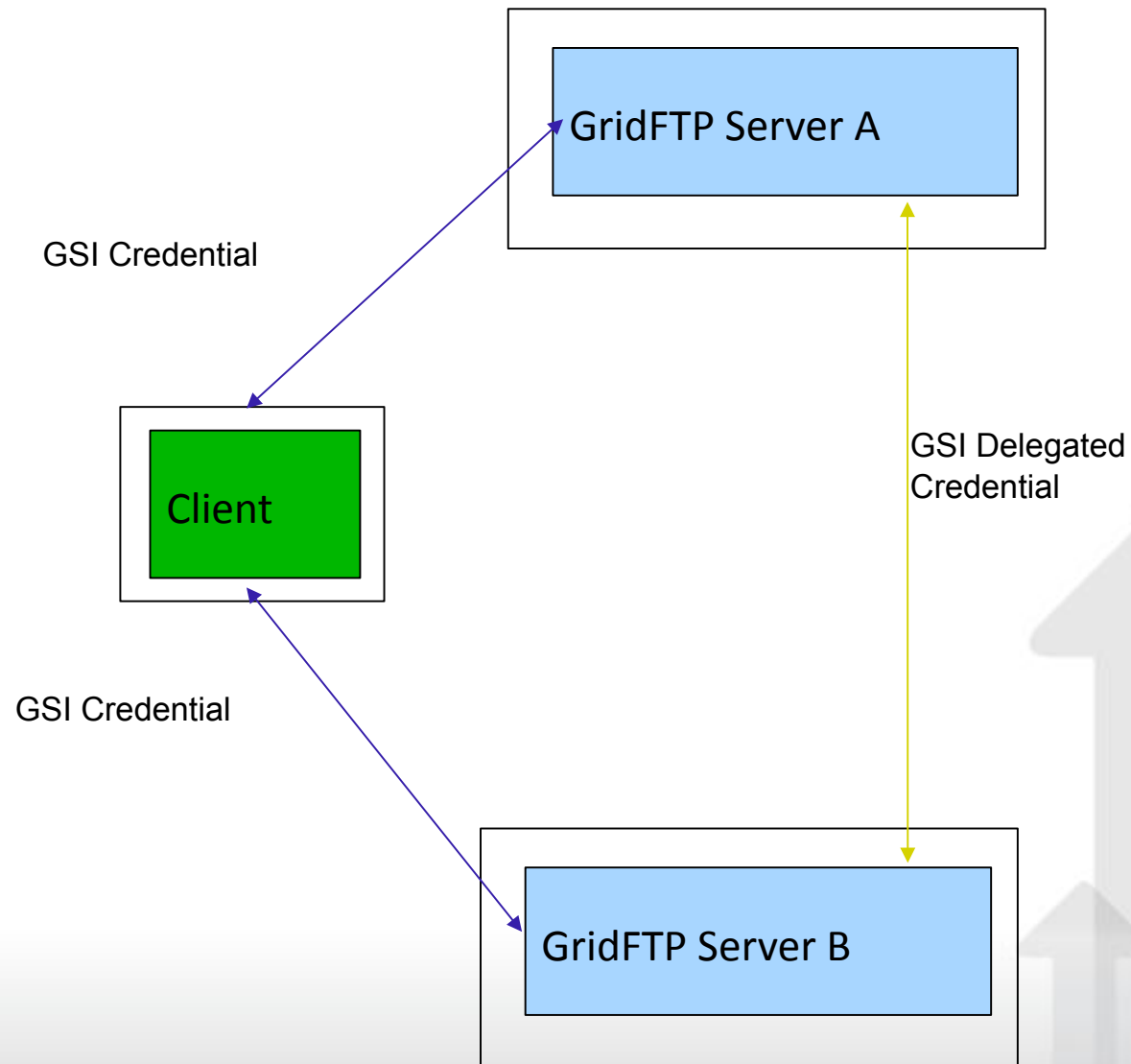
Globus Connect for Departments

- **Current Globus Connect is for individuals**

- **Many small lab and departments need Globus Connect like solution to run GridFTP with ease**

  – Single click GridFTP and Myproxy installation and setup

  – Used by all its users

  – Transfer between endpoints – neither has GridFTP server running - trivial

Questions